Beyond desire? Agency, choice, and the predictive mind

Article (Accepted Version)

BEYOND DESIRE? AGENCY, CHOICE, AND THE PREDICTIVE MIND*

Andy Clark

Department of Philosophy and Department of Informatics, University of Sussex, Brighton, UK

Abstract

'Predictive Processing' (PP) is an emerging paradigm in cognitive neuroscience that depicts the human mind as an uncertainty management system that constructs probabilistic predictions of sensory signals. Such accounts apply very naturally to perception, and have plausible extensions to motor control. But desires and motivations can seem to pose a much greater puzzle, appearing especially resistant to reconstruction by a processing story that appeals to predictions alone. I examine several versions of this worry, and show that it is fundamentally misplaced. Desires and motivations are fluently accommodated within the unifying PP schema, where they emerge as webs of prior 'beliefs' that sculpt probabilistic predictions, some of which become positioned (as we shall see) so as to bring about actions. Importantly, a single construct here plays the role of belief and desire. But what results is, perhaps surprisingly, a potentially richer landscape within which to think about agency, control, and choice.

Keywords: Desire, belief, prediction, predictive processing, agency, control, motivation

1. Introduction

Can a processing paradigm that treats biological brains as prediction engines accommodate the apparently distinctive cognitive role of wants, intentions, and desires? The problem arises because leading 'predictive processing' (henceforth, PP for short) versions of this paradigm suggest that neurally realized predictions are the only fundamental 'cognitive kind' needed to explain the full sweep of human behavior - see [Friston 2009; Hohwy 2013; Clark 2013; Clark 2016]. These neurally realized predictions may differ in their strengths, contents, and functional poise. But they are (or so it is claimed) best conceived as predictions nonetheless. In this austere landscape our actions are brought about by systemically potent predictions that act as 'self-fulfilling prophecies' [Friston 2009: 295]. We'll take a better look at how this works shortly. But at the most local (action-guiding) level, these self-fulfilling prophecies depict the act as currently taking place, and are then made true by performing the action. Predictions (of varying kinds, some more abstract and general than others)

must thus stand in for both general, long-term wants and intentions and for specific motor commands. Thus [Friston, Mattout and Kilner 2011:157] note that:

> Crucially [PP] does not invoke any 'desired consequences'. It rests only on experience-dependent learning and inference: experience induces prior expectations, which guide perceptual inference and action

In what follows, I explain, motivate, and defend this broad picture. I respond, in particular, to a raft of worries and criticisms raised in Klein [2016]. Klein's core concern is that PP, by casting everything to do with motivation, agency and desire in the form of predictions, restricts itself to an impoverished set of primitives. The result, he fears, is that the story will fail to capture key patterns in human behavior.

Another way to raise such worries (Colombo [2017]) is by pursuing the thought that predictions alone seem, on the face of it, to be motivationally inert. Motivation here signifies the whole complex of mental faculties or profiles sometimes referred to as 'conation', and taken by many to provide something that 'mere cognition' cannot. That something may be variously thought of as drive, impetus, desire, or willing. Whatever the flavor, the core idea is the same,

and was perhaps most powerfully articulated by Hume [738/2007 section 2.3.3.6] who argued that reasons without passions would be entirely unable to mandate specific actions.

Cognitions and conations, on this Humean picture, make fundamentally different kinds of contribution to the ongoing selection of action, and belief alone is insufficient for motivation. An agent might believe or even predict that her house is on fire, but what she does about that will depend on her desires. Does she desire to preserve her property or to claim on her insurance? Imagine – odd as it may seem – that she has no desires whatsoever concerning her own house. Then, faced with the opportunity to feed or to try to put out the fire, she will find no spur to act at all. Actions arise, it seems, only from the interplay between reason and passion – between belief and desire, or belief and other conative states such as fear, envy, or lust.

Similar issues arises in prominent treatments of reward- learning, such as that by Arpaly and Schroeder [2014: 286] who write that:

> Predictions are true or false; they make claims about how the world will be. Desires are neither true nor false, and they make no claims about

how the world will be.….we assume we are on safe ground in holding that the unconscious predictive system does not, itself, instantiate desires.

The distinction between beliefs and desires, and the idea that these are fundamentally different kinds of state is widely embraced in daily life. It forms a core part of common-sense (or 'folk') psychology, enabling us to say that we hope that our predictions (e.g. about feelings of coldness on the way to work in winter) do not come true. It is also deeply embedded more scientific frameworks such as statistical decision-theory (including neuro-economics and work on reinforcement learning) where it emerges as the firm separation between encodings of value or 'utility' and encodings of probability (for a useful review, see Sanfey et al [2006]).

In the rest of this paper, I'll argue that despite first appearances, the PP framework has the resources to deliver an elegant and general account of desire and motivated action. But it is an account that posits no pair of fundamentally distinct inner cognitive kinds corresponding to traditional notions of belief and desire.

Section 2 displays the basic form of the PP treatment and its extension to action. Section 3 addresses the matrix of worries raised by Klein [2016]. In section 4, I widen the discussion to include some related issues raised by Holton [2016]. These concern the proper framing of addiction and the resulting need (in PP) to invoke multiple levels and types of prediction, including (especially) predictions of the hedonic consequences of our own actions. Building on all this, Section 5 shows how PP offers a general picture of human agency. The paper ends (section 6) by asking whether the story on offer is eliminative or reconstructive regarding the folk ontology itself.

The conclusion, in a nutshell, is that the behavioral patterns we think of as reflecting desires can be successfully and plausibly conceived as underpinned by high-level predictions – ones that, in context, recruit further predictions whose proprioceptive consequences are then made true by action. This picture is revisionary insofar as it depicts our internal cognitive economy as operating using a single construct (model-based prediction) that plays the role of both beliefs and desires. But surprisingly (or so I argue) the upshot is not an impoverished vision of the human mind so much as an appreciation of a richer landscape within which to think about agency, control, and choice.

2. The Basic Schema

'Predictive processing' (PP) has been positioned as a new paradigm for understanding perception, reason, and action [Friston 2005; Hohwy 2013; Clark 2013; Clark 2016; Pezzulo, Rigoli, and Friston: 2018]. Viewed from a certain distance, PP is just a process model that extends the popular picture of the 'Bayesian Brain' (see e.g. Knill and Pouget [2004]) to include action. Bayesian brains optimally combine prior knowledge with new (sensory) evidence so as continuously to update their models of how things stand in the world. Real brains may approximate this kind of optimal updating in various ways (for some discussion, see Wiese [2016]).

PP constitutes one such family of ways, casting human brains as experts at minimizing their own long-term expected prediction error (for discussion, see [Howhy 2013:175-6; Friston 2018: 579]. Prediction error itself is simply the difference (residual error) between current predictions, reflecting the brain's best model-based guess at the evolving sensory flow, and the incoming sensory evidence itself. That error is computed, PP suggests, moment-by-moment, and at every level of processing. When error is adequately quashed, no further processing is required. But where there is mismatch, something needs to give. Crudely, we either revise our best model-based guess at how the world is, or we

alter the world or the way we are probing the world, so as to bring the sensory evidence more into line with our current predictions. The latter delivers action, the former perception, and the two constantly co-evolve as we probe and manipulate the very world we perceive (see Friston, Daunizeau et al [2010], Clark [2016], Fabry [2017], Bruineberg et al [2016], Kirchhoff [2018]).

Consider reaching for a glass. To bring about the motor action, the PP system (see Shipp et al [2013]) predicts the evolving states of muscle spindles, tendons, and joints that the reaching action demands. Since those states are not yet actual, a suite of prediction errors results. These errors are then systematically quashed by moving the body so as to make the flow of predictions come true. Cashed out by simple reflex arcs, this turns out to be an elegant and economical means of delivering motor control (Adams, Shipp, and Friston [2013]). The same basic schema applies (see section 4 below) to desired consequences at longer timescales.

Crucially, PP depicts the whole prediction-error minimizing process as nuanced and orchestrated at every level by independently coded measures of self-estimated uncertainty ('precision'). For example, when viewing a familiar scene in clear daylight, the brain may assign high reliability (high certainty, high precision) to the visual information, while on a foggy day visual information

might be assigned less reliability (less certainty, less precision) relative to stored information about the environment. Variable precision-weighting is here thought to implement attention, in all its many forms (see Feldman and Friston [2010]). So another way to think about precision-weighting is as the brain's way of stressing one body or type of information over another, and (in just the same way) balancing current top-down predictions against incoming sensory evidence – for a full discussion, see Clark [2016] chapter 2. Quite generally, then 'precision-weighting' (whose many mechanisms include neurotransmitters and time-locked neuronal oscillations) weights predictions and prediction errors according to systemic confidence in their value, context-varying usefulness, and reliability for the task at hand.

Variable precision-weighting makes these systems extremely flexible, allowing them to repeatedly redeploy stored knowledge for different purposes and in widely varying contexts. It is also the key to action. Bringing about action in this way requires attenuating (assigning low precision to) the sensory information currently indexing the *actual* disposition of the body, so as to enable precise proprioceptive predictions (corresponding to some desired trajectory) to prevail, in the manner describe above. Intentional action thus depends upon a delicate balance that combines precise proprioceptive predictions with attenuated information concerning current bodily states

(Brown et al [2013]). An interesting consequence, as noted by Wiese [2017] is that action thus depends on a kind of 'systematic misrepresentation' of how our body is currently arrayed in space!

Contrary to both the folk and standard neuro-economic images, this PP picture marks no fundamental distinction corresponding to cognitive and conative states. Instead, PP treats all forms and timescales of behavioral control as the results of precision-weighted neural predictions. These predictions issue from many different (but interacting) neural areas, and are cast at many levels of granularity, as we will later see.

PP, as even this highly truncated sketch shows, has a viable mechanism for bringing actions about. But the mere presence of such a mechanism may not be enough, if PP cannot explain when and why that mechanism becomes active. Why should we get rid of some prediction errors by action, when we might just as well (it seems) simply revise our predictions and suffer the consequences? More dramatically, it may be asked why, in the apparent absence of distinctly desire-like states aiming at life, light, and food, don't prediction agents simply find a dark corner and stay there, slowly dying of hunger and thirst while very successfully predicting that dismal but reliable sensory flux? This worry – the so-called 'Darkened Room Objection' (see Friston, Thornton,

and Clark [2012]) forms the starting point for an important recent critique of the PP proposal, to which we now turn.

3. What Predictive Coders Predict

In his [2018] paper 'What Do Predictive Coders Want?' Colin Klein presses these issues through multiple argumentative iterations. Klein's starting point is the infamous 'Darkened Room' puzzle rehearsed above. The core of that puzzle, as Klein [2018: 2554) sees it, is that "prediction alone…is not enough to get us to the adaptive actions that we in fact perform".

Let's move fairly swiftly over the first steps here. As we saw, PP has core, well-understood resources able to position certain sensory predictions so as to bring about local actions. Specifically, high precision proprioceptive predictions act as motor commands able to program the right actions when the opportunity arises. Better yet, (as we'll see in more detail in section 4 below) the same kind of story applies at longer timescales too, so that long-term plans and projects are realized as standing predictions to the effect that we will initiate project-furthering actions as and when the opportunity arises. Thus, a long-term prediction that I will pass my yacht-master exam acts as a kind of constant opportunity filter, so that when chances for practice and improvement arise,

they vie with other current opportunities to recruit actions. The winning predictions at a given moment spawn local proprioceptive predictions, and these are cashed as actions.

Importantly, precise *proprioceptive* predictions are always preferentially positioned to be quashed by action rather than by the alteration of what is predicted (for the full mechanistic story, see Shipp et al [2013]). To ensure this preferential positioning requires, as we saw, a concomitant *attenuation* of the (veridical) sensory evidence specifying the current state of the bodily plant. In order to move, we must actively downgrade (dis-attend to) information concerning the current bodily state, so as to allow the motor image of the predicted state to prevail[1]. The upshot is that "sensory attenuation is a necessary precondition for – and part of – an intended movement" [Owens et al 2018:177]. As the same authors later comment "Without this functional change in gain, prediction errors would lead to revised predictions rather than action" [Owens at al 2018:180].

[1] This is by no means an ad hoc addition to the story. Recent PP treatments of disorders of movement, such as Parkinson's disease, all point to empirically confirmed disruptions of sensory attenuation (see Brown, Adams et al [2013], Parees at al [2014] – see also Palmer et al [2016]).

PP thus possesses both a mechanism for engaging and controlling action (via descending proprioceptive) predictions *and* a systematic way to determine when to revise predictions and when to alter body (and sometimes world) by engaging action. But the real heart of Klein's worry hereabouts, I suspect, lies elsewhere. It is not the lack of a workable prediction-based means for entraining action, nor is it he need for a systematic proximal means of knowing when to act rather than revise beliefs (which is taken care of by the preferential positioning of precise proprioceptive predictions, whenever they are generated). Instead, the worry may be that PP seems (to Klein) to lack an account of *when and why* we select the specific actions we do. Translated into the PP framework, this means an account of why it is that some behavioral options get assigned high enough precision to remain 'in the driving seat' for the control of action at a certain moment in time. Thus Klein [2018:2545] asks:

> Why not, for example, eliminate the error caused by the prediction that I'll eat by revising that prediction rather than getting some food?

There is, to repeat, a very effective *proximal* mechanism for resolving this. If a food-seeking action such as picking up the phone to order a takeaway, is assigned high precision, current sensory information specifying that my hand is not currently moving is attenuated, and the food ordering scenario prevails. But

the real question then becomes: why are the checks and balances in the PP economy thus and so? Why is the strong prediction of food-ordering (that programs the rest of the economy) currently entrenched?

One response at this point has been to depict some predictions as simply more deeply ingrained than others. Perhaps we cannot help (at some level) but predict food and safety, no matter what the weight of lifetime evidence - from, for example, living in a war zone – suggests. This fits with the picture presented by Friston [2011] according to which some predictions are in effect simply definitional of the creatures concerned. As an evolved being, I thus chronically predict a certain temperature range – the one needed for a creature like me to survive. Predictions are made on the basis of a generative model, and the generative model that we (considered as whole embodied organisms) instantiate will have been shaped by both evolution and lifetime learning so as to be one that ensures we are deeply disposed to predict, with high, action-entraining precision, the kinds of sensory state that help keep us alive and viable. Among such deep-set predictions we will find, for example, ones that mandate keeping key features of the bodily plant within tolerable limits. Deep-set interoceptive predictions [Seth 2015] may thus ensure that darkened food-free rooms hold no allure for the normally functioning human agent.

But of course, such deep-set predictions get us only so far. At some point, the PP theorist needs to accommodate the ordinary shifting webs of (as we would ordinarily say) desire: the ebbs and flows of intention that sometimes lead us to play the piano, then to work on a paper, then to order a Chinese rather than an Indian takeaway, watch a certain movie, and so on.

This is where the shifting web of precision assignments, and the multiple time-scales of prediction, must work together to realize the full spectrum of context-varying motivation and choice. We select specific actions, PP claims, because we already (both personally and sub-personally) encode a host of (complexly interacting) multi-timescale predictions, and because our brains constantly adjust the precisions with which these predictions are held. As both our inner states (hunger, thirst etc.) and outer contexts ebb and flow, some predictions enjoy increased precision, becoming positioned to drive immediate actions while others remain in the background, awaiting the right opportunity to arise. Right now, for example, it is my high-precision prediction that I am exploring Klein's argument that is selecting my actions – both at the level of looking up various papers to check my claims, and then making specific key-strokes (cashing out precise proprioceptive predictions) on my computer.

The skeptic might press the issue by asking why a given agent predicts the very things she does, with their various weightings. Perhaps she chooses to order tofu rather than chicken for the take-away. Why did her lifetime learning position the tofu prediction so as to trump her colleague's suggestion of chicken? The PP mechanism itself offers no concrete story here. But neither does an account that appeals to separately encoded motivations or desires. The question of why certain desires dominate action and choice at time T is now simply replaced with the question of why certain predictions are assigned high precision at time T. Neither one of these questions is (as far as I can tell) any easier, or simpler, than the other.

Some commonplace human experiences may seem to work against this suggested assimilation of altered desires and motivations to altered precision-weighted predictions. Most notably there is the feeling that, as Klein [2018] puts it, *something else* needs to be there to ensure we keep moving and acting rather than simply *tracking* states of our own body and the wider world. Predictions just do not seem like the kinds of things that have the requisite 'oomph' to drive us to action – they seem, on their own, to lack intrinsic motivating force. But this, as we saw earlier, is not the case. The motivating force is instead provided by the shifting assignments of precision to varying

predictions. Highly-weighted predictions drive actions, and do so every bit as directly as (on more standard models) would strong preferences and desires.

The devout Humean, confronted with the this proposal, may insist there must still be something (some kind of inner or mental state) that is not itself cast in terms of predictions that sets it all in motion: some additional inner force favouring some actions or outcomes over others. To really get to grips with the PP story, it is important to stress that this is not the case. Instead, there is simply a rich precision-inflected predictive economy some of whose key aspects have been put in place by evolution, others by lifetime learning. That web of predictions changes and evolves as new contexts arise and new information is acquired, by means of ongoing self-organization around the organismically accessible quantity of prediction error. It is our changing predictions (both responding to, and inflected by, changing assignments of precision) that must now explain all the kinds of changes in response ordinarily captured by folk-psychological talk of changing motives, goals, and desires.

Superficially still puzzling, perhaps, is the nature and origin of apparently idiosyncratic individual desires and motivations. Consider, once again, my desire to go to see a certain movie tonight. PP realizes this desire as a high-level prediction that (when estimated as sufficiently precise) entrains apt actions at

many time-scales. This is actually a familiar (Bayesian) trick for transforming control and selection problems into belief-based inference problems, enabling that well-understood apparatus to deliver choice and planning behavior too (see, for example, [Toussaint 2009;Todorov 2009]). But where did that high-level prediction itself come from? The answer is not obvious – but nor (to repeat) is it any more obvious if we stick with unanalyzed appeals to motivation or desire. Either way, what really needs to be explained is the process of change in long-term states that bring about actions. As PP agents move, think, and act in their worlds they change and alter their own webs of precision-inflected multi-level probabilistic belief. In that way they constantly alter the processing that in the future brings about their own actions.

Klein remains skeptical, and for two main reasons. The first is that all we may have then done, he argues, is recreate the classic belief/desire split using PP resources. In one way this is absolutely right – the PP story now accommodates all the behaviors we ordinarily use talk of beliefs and desires to capture. But PP aims squarely at a sub-personal mechanism. And at that level, PP has indeed replaced the traditional constructs of beliefs, desires, and rewards with a single construct – that of predictions which, when they are held with high precision and have proprioceptive consequences, entrain apt action. I return to this issue in Section 5.

Klein's second worry is that these interacting predictions will now ignite an explosion of complexity. For example, how do I rank which action gets high precision next? Klein [2018:2550] argues that "ordinary belief-desire models can avail themselves of the standard combinatorial resources of computational theories to try to sort out these problems". Here, Klein either greatly over-estimates the power and successes of the standard stories or greatly under-estimates the resources of the PP alternative. There is no doubting that selecting which action to perform next, given a large body of world-knowledge, is computationally challenging. But it is no more challenging using PP resources than it is using more traditional ones. All the results of reinforcement learning (to take just one example) are available to PP, simply by re-casting value and reward in terms of precision-weighted prediction error (surprisal) minimization. Instead of aiming sub-personally at rewards, PP agents aim sub-personally at minimizing surprise about future states. This, in turn, requires them to adopt policies (long-term behavioral strategies) that will deliver just that (for the full story, see [Schwartenbeck et al 2013; Pezzulo, Rigoli, and Friston 2018]). Moreover, neither the classic story nor the PP alternative has very much to say about the origins of idiosyncratic desires (idiosyncratic predictions, for PP), save that organisms start somewhere, then move around

their physical and social worlds in ways that progressively install their tendencies to seek out science fiction, or horror stories, or to breed exotic fish.

One final point deserves a brief mention. Klein also worries, towards the end of the 2018 paper, that PP inherits (from the so-called 'free energy principle', which is not our focus today) a kind of objectionable absolutism. The worry is that if the goal is really to minimize prediction error (free energy) then only actions that do so will be selected, leaving no room for actions that reduce but do not quite minimize that same quantity. This is simply mistaken. The best way to minimize expected prediction error over time may well be to perform actions that have exactly the 'partial' character that Klein celebrates. Eating half an apple is better than eating hot lava (to use his own example) because eating half the apple plausibly reduces some prediction error whereas eating the lava actually increases it. Over time, selecting actions that make partial headway with goals serves to minimize overall expected error (see [Pezzulo, Rigoli et al 2015, 2018; Friston, Schwartenbeck et al 2014]). Perhaps relatedly, there is a cost-of-modeling issue that Klein has failed to spot. PP aims [Clark 2016: chapter 8] at the use of the most minimally expensive (fewest parameter) models that will reduce the greatest amount of expected prediction error. This immediately implies that solutions that are intuitively 'just-good-enough' are often PP-optimal.

4. When Predictions and Desires Conflict

There is a related issue nicely foregrounded in Holton [2016]. It concerns the frequent appearance of conflict between our best predictions and our strongest desires. Imagine you have an itch. You want to scratch it, but you predict that this would be bad for your skin. Still, the urge is overwhelming. You scratch and suffer the consequences. How are we to understand this kind of daily misadventure? According to the standard picture, the conflict arose between two fundamentally different 'mental kinds' – what we want (to scratch the itch) and what we predict (that this will be bad for our skin). And in this case, the desire won out.

PP tells a different story. If PP is correct, the weak-willed scratcher harbors conflicting belief-like states. These states are really predictions[2], some of which (those assigned high precision) get to bring about actions. They include both the prediction that scratching will harm the skin and should be avoided, and the prediction that I am going to scratch that itch here and now. The latter, having proprioceptive consequences that are currently predicted

[2] More accurately, these comprise both probabilistic priors (when they are long-term or standing states) and predictions (when they are active states).

with high precision, renders the former transiently impotent. Scratching ensues. As always, PP here replaces desires with predictions, which select gross actions only when they have proprioceptive consequences that are being assigned high precision. In one sense, these predictions are beliefs about our own future behaviors. But high-level predictions (when opportunities are spotted, so they spawn precise proprioceptive predictions) are also the drivers of the predicted behavior. This means they are best not assimilated to *either* the folk-psychological category of desire *or* the folk-psychological category of belief.

Holton [2016:10] suggests that assimilating desires to predictions "doesn't do justice to the multiplicity and malleability of human desire", citing cases in which drug addicts (for example) may seek out a desired substance while simultaneously believing that taking the drug won't bring them either happiness or pleasure. Instead the drugs are said to be "simply intrinsically wanted". This fact is then claimed to be "radically at odds with anything the predictive processing account says about us". But we can now see why this need not be the case. For what looks, from Holton's perspective, to be a clear case of conflict between belief and desire may now be re-cast as a difference between predictions of different kinds. Specifically, the predictive processing story firmly distinguishes [Friston, Shiner et al. 2012] between action-entraining high-precision predictions concerning what I will do and predictions of the

hedonic (interoceptive) outcomes of those very actions. PP thus accommodates the fact, highlighted by Holton, that drug users often do not believe/predict that taking the drugs will actually lead to happiness. But what they do powerfully (if often sub-personally) predict is seeking and ingesting the drug. PP thus fluidly reconstructs the useful distinction between 'wanting' and 'liking' suggested by Berridge [2007]. For a full PP treatment of this, see Schwartenbeck, Fitzgerald et al [2015].

More generally, and given that the addict need not predict that the drugs will bring pleasure, PP remains well-placed to explore a wide variety of promising accounts in which many of the addict's experiences and actions are the results of interacting sub-personal (non-conscious) predictions and expectations. For example, it has been known since the work of Siegel [1983] that mere exposure to the paraphernalia associated with drug use can trigger physiological symptoms of withdrawal. Their explanation appeals to a physiological preparation for the predicted effects of the drug (inducing so-called 'tolerance'). Thus Siegel and Ramos [2002:171] comment that "some drug 'withdrawal' symptoms are more accurately drug 'preparation symptoms' ". One very natural way to think about such results is to see the setting and paraphernalia as triggering a sub-personal stream of both inward-looking (bodily) and outward-looking predictions, which actively warp sensation in the

direction of certain strongly predicted effects. Such effects, PP suggests, are mediated by dopamine and other neurotransmitters that modulate the delicate balance between sensory evidence and 'top-down' predictions, and now fall neatly into place within a larger framework that is being successfully applied to a wide range of medical symptoms (see e.g. [Bergh et al 2017]).

All this suggests that there is nothing austere or conceptually impoverished about the PP story concerning motivated behavior. On the contrary, where more traditional stories posit a single division between belief-like and desire-like states, PP depicts as many varieties of 'predictive controller' as there are varieties of ways and means of contextualizing lower-level response. Fears that PP is fated to miss complexity or merely reconstruct the traditional story are deeply misplaced. Instead, the new story is potentially more powerful, recognizing a richer and more subtly varied continuum in which deep goal hierarchies constantly inflect perception and action in ways determined by ever-shifting webs of precision. In the remaining two sections, I expand further upon this claim, and discuss the extent to which PP is revisionary with regard to our normal understanding.

5. The Stepladder to Agency

To recap, PP offers a promising and mechanistically viable account of simple intentional motor actions. Precise high-level predictions conspire to enable motor 'wishes' – here realized by precise proprioceptive predictions combined with attenuated current sensory information - to entrain the bodily plant. This works without any need to introduce distinctly conative factors. That same story must apply 'all the way up'. Our action-guiding proprioceptive predictions are themselves caused by even higher-level and longer time-scale predictions – predictions about our own future behaviors and our own resulting future states. These now form a kind of temporal stepladder in which different beliefs about our own future actions entrain those very actions by triggering apt proprioceptive predictions (the proximal action-makers) when good opportunities arise. In this way (laid out in more detail by [Pezzulo et al 2015; 2018]) nested beliefs entrain actions at many interacting time-scales by bringing about predicted sensory flows.

As a cameo, suppose that as I move around my world today, harvesting sensory information, I come to believe/predict that I will meet you at the movie-theatre for the 8 pm showing tonight. This high-level belief yields activity (perhaps I consult the web) that in turn leads me to believe/predict that I will get the 730

bus. That prediction then acts as a kind of (defeasible) mini-policy that enslaves further apt motor action when it is time for me to leave the house. What we do is determined, this story suggests, by precise (highly weighted) high-level predictions that respond to worldly opportunities by delivering a whole swathe of apt lower-level sensory and motor predictions. High-level predictions thus act as controllers, determining how we act in the world. Such predictions are both somewhat belief-like (being about what is predicted to occur) and somewhat desire-like (being functionally poised to bring those very things about).

These are fully-fledged exercises of agency, and they involve what has sometimes been described [Pezzulo et al 2015; Seth 2015] as 'counterfactual prediction'. In such cases a temporally deep [Friston, Rosch et al 2017; Friston 2018] prediction-issuing generative model includes predictions of what we *would* experience (e.g. arriving at the movie theatre) *if* we acted in some specific way. For this to be possible, advanced agents must command a probabilistic model of how their own sensory experience would be altered or updated (and expected prediction error reduced) were they to perform such and such actions. The active human agent thus turns up, implicitly, as a kind of latent variable (a so-called "hidden cause") in her own model of future sensory unfoldings. But crucially, as argued by Friston and Adams [2012], she is a hidden cause of

sensory unfoldings that are highly controllable by her own actions – unlike many other hidden causes, such as weather and volcanoes. In this way, agency itself emerges naturally from prediction machinery operating over multiple time-scales.

At this point new opportunities arise. Future goals can now be approached (as in the movie-going case) by the adoption of sensible long-term policies, where these are simply belief-complexes that entrain sequences of actions, each of which delivers predictable sensory flows that bring us closer to some goal. We here enter the space of long-term goal-directed systems, ones that:

> ….represent counterfactual future states, and minimize the difference between the preferred or goal state and outcomes predicted from the current stat…[a]…prospective form of control…supported by the ability of higher hierarchical levels to anticipate the future and to select policies that enslave action. [Pezzulo et al. 2015: 24]

Ongoing arbitration between multiple goals is here determined by balancing standing precision (reporting long-term value to the agent) and perceived opportunity. The latter is cashed [Friston, Schwartenbeck, et al. 2014] by

currently attainable, sufficiently valued sub-goals being accorded extra-high precision, thus preferentially entraining action. Dopamine – a key player in the game of precision-weighting – here plays a major role reporting what is both salient and actionable in the present web of opportunities. In this way stories that appeal to distinct cognitive and conative factors, such as the neuroeconomic models described by Sanfey et al [2006], are fully subsumed under the more encompassing and unified picture of an economy of predictions nuanced and enabled by shifting matrixes of precision-weighting.

6. Beyond Belief, Beyond Desire.

I have argued that, using the resources sketched above, PP is able to capture whatever complex webs of changing desire and motivation the standard story posits. But this suggests another kind of worry. At this point, we may seem simply to have *reconstructed* the standard story, but to have done so in a way that loses transparency without adding any real value. For it requires us to posit complex webs of (perhaps mutually inconsistent) predictions, some playing the role of genuine world-reflecting beliefs, others that of action entraining predictions.

Klein [2018:2551] presses just such a concern, worrying that a sufficiently powerful PP story of this kind closely resembles traditional motivation-based accounts but that:

> …the bookkeeping with one state gets complicated, while a system with more primitives (such as a belief-desire model) can more easily keep track of shifting needs, goals, and facts about the world.

Having thus gone to some pains to show that the PP approach has all the resources to capture commonsense intuitions about behavior, we must now ask what, if anything, is really *different* enough in the PP-treatment to recommend it.

What's most importantly different is that there's now a different and arguably much more unified internal architecture, trading in a single currency (predictions, prediction errors, and their precisions). The shape of this architecture is reasonably straightforward. On top of simple reflex arcs (and autonomic homeostasis) there emerge more 'Pavlovian' controllers, whereby the sound of a ringing bell (say) comes to predict and hence bring about the sequences of interoceptive signals that normally accompany the sight and

ingestion of food[3]. On top of those emerge further 'instrumental controllers', so that if (for example) we learn that the bell rings whenever our hand moves a certain way, we may start to 'expect' to generate food or rewards by moving our hand in that fashion. Further up the ladder, as we saw, human agents, immersed in their idiosyncratic social and cultural environments, start to form and exploit 'counterfactual predictions' tracking what we would experience and infer about the world if we acted in some specific way.

What emerges is indeed a rich and unified architecture marked by the successive contextualization of control by higher and higher level predictions, implicating different neural areas and neuronal populations, but all co-operating within a single processing regime defined using a common currency of predictions, precision estimations, and prediction errors. At the bottom of the stack are simple peripheral reflexes (e.g. involving proprioceptive predictions that determine set points for stretch receptors that then automatically translate into movements). Towards the top lie more intuitively agency-reflecting (indeed, agency-constituting) predictions, such as the prediction that I will go and see such-and- such a movie tonight.

[3] Here, and in the rest of this paragraph, I briefly rehearse the much richer and more detailed picture presented in [Pezzulo et al 2015;2018]

Importantly, the overarching schema is just the same as the schema for basic perceptual inference. Here too (just as in ordinary perception) low-level predictions are contextualized by higher-and higher level predictions, nuanced by shifting precision estimations. Reflexive, habitual, locally goal-directed, and fully 'prospective' (future-oriented, in the sense of Seligman [2013]) behaviors are revealed as simply different contextualizations within a single continuum of prediction-based control.

In deploying this complex hierarchy, all the usual costs and benefits apply. Time-pressured or resource-pressured response will recruit more habitual forms of control, requiring less processing and delivering faster results (see e.g. [Clark 2016; chapter 8]. By self-organizing around changing prediction error signals, these systems generate a flux of precision estimations that allow behaviors to be controlled by just about any possible combination of experientially 'deliberative' and more 'automatic' processes. This delivers a vast spectrum of task- and context-sensitive response that potentially affords far richer possibilities than more traditional ('dual route') models of the kind effectively critiqued by Hommel and Weirs [2017].

So does the PP story really do away with motivation and desire? We need not (and should not) gloss the complex, multi-level PP story as

eliminativist in any strong sense. The patterns it depicts in human behavior are real, and the discourse 'earns its keep' (for an extended argument to this effect, see Dewhurst [2017]). But desires, intentions, and motivations are now all realized as varying forms and time-scales of prediction. Desires thus realized are standing (generative model-based) predictions that become active and positioned to entrain actions according to the context-varying flux of precision-weighting. Such a picture is revisionary with regard to the traditional picture of minds as mechanisms that work by combining instantiations of the distinct cognitive kinds of belief and desire. But the PP story is thus every bit as revisionary about beliefs (considered as parts of the machinery of mind and action) as it is about desires. For all that is posited are multi-level, multi-timescale webs of probabilistic priors that sculpt predictions, that can act in ways that are both belief-like and desire-like.

Summing up, active predictions that are accorded high precision, if they also imply precise (highly-weighted) proprioceptive consequences, get to entrain local action. These active states are self-fulfilling prophecies, capable of helping to bring about the very states of affairs they describe. Such states must arise and dissolve, self-organizing around prediction error, in ways that realize

our experiences of belief, desire, and all the other states (hope, fear, love, anger) that together make up the phenomenological flux[4].

7. Conclusions: Predictions as Controllers

Critics have questioned whether predictive processing (PP) has the resources to accommodate the full complexities of motivated human behavior. I have tried to show that such skepticism is unwarranted. Most fundamentally, this is because it is a mistake is to treat prediction as some kind of passive projection of statistics from past experience (individual or species) into the future. Instead, a major role of prediction (in PP) is to *bring about* the statistical patterns in behavior that define us both as individuals and as a species. Relatedly, the estimated precision of specific predictions is not, as is sometimes thought, simply a measure of systemic confidence in their accuracy or reliability – rather, it is a device that actively positions some predictions for the control of behavior. Although PP posits only a few types of computational states (predictions, precisions, and prediction errors), those core states can thus be functionally poised in many ways, and associated with a wide variety of cognitive contents.

[4] I have not attempted to outline the PP account of other human attitudes here, but for some hints, see Joffily and Coricelli [2013]

Neurally realized predictions here play a dual role, both responding to past experience and sculpting future choice and action. Such states are somewhat belief-like, consisting in precision-weighted predictions, but somewhat desire-like too, since they may select and entrain actions at multiple time-scales. It is perhaps surprising to see the same computational construct performing both these functions. But this allows the brain to construct motivation, motor control, and action selection using exactly the same computational palette. The result is a fluid and highly context-sensitive regime that integrates control and motivation at every level. This may slowly reveal a new and richer landscape within which to think about agency, control, and choice.

* Heartfelt thanks to the three anonymous referees whose careful and constructive comments have hugely improved this treatment. Thanks also to Anil Seth, Jakob Hohwy, Colin Klein, and Karl Friston for useful discussion of many of these ideas. The paper was written thanks to support from ERC Advanced Grant XSPECT - DLV-692739.

References

Adams R, Shipp S, and Friston K. 2013. Predictions not commands, *Brain Structure and Function* 218:3:611-43.

Arpaly, N., and Schroeder, T. 2014 *In Praise of Desire*, Oxford: Oxford University Press.

Bergh, O. Van Den, Witthöft, M., Petersen, S., & Brown, R. J. 2017. Symptoms and the body : Taking the inferential leap, *Neuroscience and Biobehavioral Reviews* 74:A:185–203.

Berridge K.C. 2007 The debate over dopamine's role in reward, *Psychopharmacology* 191:3:391–431.

Brown, H., Adams, R. A., Parees, I., Edwards, M., and Friston, K. 2013. Active inference, sensory attenuation and illusions, *Cognitive Processing* 14:4: 411-427.

Bruineberg, J., Kiverstein, J., & Rietveld, E. 2018. The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective, *Synthese* 195:6:2417–2444.

Churchland, P.M. 2012 *Plato's Camera: How the Physical Brain Captures a Landscape of Abstract Universals*. Cambridge, MA: MIT Press.

Clark, A. 2013 Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science *Behavioral and Brain Sciences* 36:3:181-204.

Clark, A. 2016 *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*, New York, Oxford University Press.

Colombo, M. 2017. Social motivation in computational neuroscience. J. Kiverstein (Ed.) *Routledge Handbook of Philosophy of the Social Mind*. New York: Routledge. 320-340.

Dewhurst, J. 2017. Folk Psychology and the Bayesian Brain. T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing* Frankfurt am Main: MIND Group. 9: 1-13.

Fabry, R. E. 2017. Predictive processing and cognitive development. In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing*. Frankfurt am

Main: MIND Group. 13: 1-18.

Feldman H and Friston K. 2010. Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience* 2:4:215:1-23.

Friston K. 2005. A theory of cortical responses. *Philosophical Transactions of the Royal Society London B Biological Sciences* 29:360:815-36.

Friston K. 2009. The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences* 13:7:293–301.

Friston K 2011 Embodied Inference. In W. Tschacher and C. Bergomi (eds) *The Implications of Embodiment (Cognition and Communication)* Exeter, Imprint Academic 89-125.

Friston, K. 2018. Am I Self-Conscious? (Or Does Self-Organization Entail Self-Consciousness?) *Frontiers in Psychology* 9:4:579.

Friston, K., & Adams, R. 2012. Perceptions as hypotheses: saccades as experiments. *Frontiers in Psychology*, 3:5:151.

Friston K, Daunizeau J, Kilner J, and Kiebel SJ. 2010. Action and behavior: a free-energy formulation, *Biological Cybernetics* 102:3:227-260.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. 2016 Active Inference and Learning, *Neuroscience and Biobehavioral Reviews*. 68:9:862-79

Friston K, Mattout J, and Kilner J. 2011 Action understanding and active inference, *Biological Cybernetics* 104:1-2:137–160.

Friston, K. J., Rosch, R., Parr, T., Price, C., and Bowman, H. 2017. Deep temporal models and active inference, *Neuroscience and Biobehavioral Reviews* 77:6: 388–402.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., and Dolan, R. 2014. The anatomy of choice, *Philosophical Transactions of the Royal Society London B Biological Sciences* 369:November: 1655

Friston K, Shiner T, FitzGerald T, Galea JM, Adams R, et al. 2012 Dopamine, Affordance and Active Inference, *PLoS (Public Library of Science) Computational Biology* 8:1: e1002327.

Friston, K., Thornton, C., and Clark, A. 2012. Free-Energy Minimization and the Dark-Room Problem, *Frontiers in Psychology* 3:130: 1–7.

Hohwy, J. 2013. *The Predictive Mind*, New York, Oxford University Press.

Holton, R. 2016. Review of Surfing Uncertainty, *Times Literary Supplement* October 7: 10-11.

Hommel, B and Wiers, R. 2017. Towards a Unitary Approach to Human Action Control, *Trends in Cognitive Sciences* 21:12:940-949.

Hume, D 1738/2007 *A Treatise of Human Nature: A Critical Edition*, David Fate Norton and Mary J. Norton (eds.), Oxford, Clarendon Press.

Joffily, M., & Coricelli, G. 2013. Emotional Valence and the Free-Energy Principle. *PLoS (Public Library of Science) Computational Biology* 9:6: e1003094.

Kirchhoff, M. 2018. Predictive processing, perceiving and imagining *Philosophical Studies* 175:3:751–67.

Klein, C. 2018. What do predictive coders want? *Synthese* 195:6:2541–2557.

Owens, A.P., Allen, M., Ondobaka, S., Friston, K.J., 2018. Interoceptive inference: from computational neuroscience to clinic, *Neuroscience and Biobehavioral Reviews* 90: July: 174-183.

Palmer, C., Davare, M. and Kilner, J. 2016 Physiological and Perceptual Sensory Attenuation Have Different Underlying Neurophysiological Correlates, *Journal of Neuroscience* 36:42:10803-12.

Parees, I., Brown, H., Nuruki, A., Adams, R.A., Davare, M., and Bhatia, K.P., 2014. Loss of sensory attenuation in patients with functional (psychogenic) movement disorders, *Brain* 137:11:2916–21.

Pezzulo, G., Rigoli, F., & Friston, K. 2015. Active Inference, homeostatic regulation and adaptive behavioral control, *Progress in Neurobiology*, 134:November:17–35.

Pezzulo G., Rigoli, F. Friston, K. 2018. Hierarchical Active Inference: a Theory of Motivated Control, *Trends in Cognitive Sciences* 22:4: 294-306.

Sanfey, A. G., Loewenstein, G., McClure, S. M., & Cohen, J. D. 2006. Neuroeconomics: Cross-currents in research on decision-making, *Trends in Cognitive Sciences*, 10:3:108–16.

Schwartenbeck, P., FitzGerald, T. H. B., Mathys, C., Dolan, R., Wurst, F., Kronbichler, M., & Friston, K. 2015. Optimal inference with suboptimal models: Addiction and active Bayesian inference, *Medical Hypotheses*, 84:2:109-17.

Seligman, M. E. P., Railton, P., Baumeister, R. F., & Sripada, C. 2013. Navigating Into the Future or Driven by the Past? *Perspectives on Psychological Science*, 8(2), 119–141.

Seth, A. K. 2015. The Cybernetic Bayesian Brain, in T. Metzinger & J. M. Windt (Eds). *Open MIND*: 35(T). Frankfurt am Main: MIND Group. 1-24.

Shipp, S., Adams, R. a, & Friston, K. J. 2013. Reflections on agranular architecture: predictive coding in the motor cortex, *Trends in Neurosciences*, 36:12:706–16.

Siegel, S. 1983. Classical Conditioning, Drug Tolerance, and Drug Dependence, *Research Advances in Alcohol and Drug Problems*: Volume 7. Boston, MA: Springer US. 207–246.

Siegel, S., & Ramos, B.C. 2002. Applying laboratory research: Drug anticipation and the treatment of drug addiction, *Experimental and Clinical Psychopharmacology* 10:3:162–183.

Todorov, E. 2009. Efficient computation of optimal actions, *Proceedings of the National Acadamy of Science USA* 106:28:11478–83.

Toussaint, M. 2009. Probabilistic inference as a model of planned behavior, *Künstliche Intelligenz* 3:9:23-29.

Wiese, W. 2016. What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences*. 6 :4:715-36.

Wiese, W. 2017. Action Is Enabled by Systematic Misrepresentations *Erkenntnis* 82:6:1233–52.